正則化法を用いたロジスティック回帰モデルによる多次元データでの変数選択手法に関する研究 (Statistical Experiment and Its Related Topics)

AUTHOR(S):

阪本, 亘; 高橋, 史朗; 竹内, 正弘

CITATION:

阪本, 亘 ...[et al]. 正則化法を用いたロジスティック回帰モデルによる多次元データでの変数選択手法に関する研究 (Statistical Experiment and Its Related Topics). 数理解析研究所講究録 2010, 1703: 32-52

ISSUE DATE:

2010-08

URL:

http://hdl.handle.net/2433/170026

数理解析研究所講究録
第1703巻 2010年 32-52

正則化法を用いたロジスティック回帰モデルによる多次元データでの変数選択手法に関する研究

北里大学大学院薬学研究科 臨床医学(臨床統計学)

Department of Clinical Medicine (Biostatistics), Graduate School of Pharmaceutical Sciences, Kitasato University

阪本 亘(Wataru Sakamoto)
高橋 史朗(Fumiaki Takahashi)
竹内 正弘(Masahiro Takeuchi)

1.序論

遺伝子領域における著しい科学技術の進歩により登場したマイクロアレイを用いることで,1 度に数万にも及ぶ遺伝子の発現情報を調べることが可能となった. そして,得られてくる遺伝子発現情報と疾病の状態や薬剤による副作用の有無といった臨床情報との間の関連性を解析し,患者の予後を規定している遺伝子あるいは遺伝子群を探索することが試みられている. これにより,新たな創薬ターゲットを発見することや遺伝子発現情報から患者個々の予後を精度良く予測しテーラーメイド医療に応用することなどができるのではないかと考えられており,マイクロアレイを用いた臨床研究が盛んに行われている.

マイクロアレイデータ解析における統計学の大きな課題の 1 つは,症例数(n)に比べて遺伝子数,すなわち,独立変数の数(p)が非常に大きい"$p \gg n$"の問題である. 遺伝子データを扱わない医学研究では"$n > p$"の状況が一般的であり,例えば,応答変数が連続値の場合には最小二乗法を用いた重回帰分析,2 値の場合には最尤法を用いたロジスティック回帰分析が広く用いられている. しかし,"$p \gg n$"の状況下において最尤法は解の不定問題に直面してしまう. この問題に対処できる統計手法として,回帰係数の L1 ノルムや L2 ノルムを罰則項に利用した回帰分析手法である Lasso (Tibshirani (1996))[1]や Elastic net(Zou et al. (2005))[2]が近年注目を浴びている. 両手法では,縮小推定により多くの回帰係数の推定値を正確に 0 にできるという好ましい特徴がある. したがって,得られてくるモデルは少数の独立変数から成る簡素なモデルであり,自動変数選択が可能となる.

しかし,これらの手法を遺伝子データへ応用することを考えた時,独立変数間の相関が新たな問題として生じる. 遺伝子間には高い相関があることが知られており,それらの遺伝子はグループを形成していると考えることができる. この場合,Lasso では応答変数と関連のある遺伝子グループの中から 1 つの遺伝子のみをランダムに選択

してくる性質がある．つまり，真に関連のある遺伝子を選択し損なう可能性が増大するという欠点が生じてしまう(“真陽性”の減少)．一方，Elastic net では相関が高い遺伝子をグループとして選択できるので，Lasso の欠点を克服できるものの，応答変数と関連のない遺伝子を多くモデルに選択してきてしまう可能性が高くなるという欠点が生じてしまう(“偽陽性”の増大)．ここで，“真陽性”とは，本当は応答変数と関連のある独立変数のうち，回帰係数の推定値が 0 ではない独立変数の数を意味し，また，“偽陽性”とは，本当は応答変数と関連のない独立変数のうち，回帰係数の推定値が 0 ではない独立変数の数を意味するものとする．

そこで，Elastic net の L1 ノルムの部分に独立変数毎に異なる重みを加えた罰則項を用いて，任意の繰り返し数 M の範囲においてモデル評価基準が良くなり続ける限り，回帰係数の推定と重みを更新していく，Recursive elastic net (Shimamura et al. (2009)) [6]という新たな回帰分析手法が提案された．この手法により，相関を考慮できるという Elastic net の好ましい性質を保持しつつ，弱点である“偽陽性”を大きく減らせることが，正規線形モデルとベクトル自己回帰モデルの枠組みにおいて示された．しかし，応答変数が 2 値や生存時間の場合には，その性能は未知である．したがって，変数選択の点からみると，独立変数間に高い相関を持つマイクロアレイデータを解析する際，Lasso や Elastic net に比べ，より妥当な手法と思われる Recursive elastic net のもつ(1)独立変数毎の相対的重要性を考慮できる，(2)繰り返しの推定を行う，という 2 つのアイデアを他の回帰モデルに適用し，その性質を調べておくことは有用と思われる．

そこで，本研究では，シミュレーションとマイクロアレイの実データ解析を通して，上に述べた Recursive elastic net の 2 つの考え方をロジスティック回帰モデルに導入した手法(以下，検討手法)と，Lasso, Elastic net を変数選択の観点から比較検討することを目的とした．

第 2 章では，本研究と関連する L1 ノルムや L2 ノルムを用いた回帰分析手法について，“$p \gg n$”での変数選択における性質を中心にレビューを行う．第 3 章では，検討手法と Lasso, Elastic net をモンテカルロ・シミュレーションにより比較する．第 4 章では，実際のマイクロアレイデータに対して各手法を適用した結果を示す．そして，最終的な考察を第 5 章で行う．

2.正則化法を利用した回帰分析手法

以下，症例数を n，独立変数の数を p とし，Y を 0 または 1 を取る 2 値応答変数，独立変数を $X_1, X_2, \cdots\cdots, X_p$ と表すことにする．λ_1 及び λ_2 は正則化パラメータと呼ばれるものであり，縮小の程度を調節する役割を果たしている．最適な値は(0, ∞)の範

囲において，交差検証法や情報量基準を用いることで決定される．

2-1.最尤推定

正則化法への導入として，医学研究において広く利用されているロジスティック回帰モデルと最尤法について説明する．ただし，この項でのみ“$n>p$”を仮定している．

第 i 症例の独立変数 $\boldsymbol{X}_i=(1, X_{i1}, X_{i2}, \cdots\cdots, X_{ip})^{\mathrm{t}}$ を与えた時の成功確率 $\Pr(Y_i=1\mid \boldsymbol{X}_i)$ を π_i とおくとき，ロジスティック回帰モデルは以下のように表わされる；

$$logit(\pi_i)=log\left(\frac{\pi_i}{1-\pi_i}\right)=\boldsymbol{X}_i^{\mathrm{t}}\boldsymbol{\beta}$$
$$\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots\cdots,\beta_p)^{\mathrm{t}}$$

上式における回帰係数 $\boldsymbol{\beta}$ の最尤推定値 $\hat{\boldsymbol{\beta}}_{MLE}$ は最尤法により求まり，これは経験ロス関数である負の対数尤度関数 $-l(\boldsymbol{\beta})$ を最小にする解である；

$$\hat{\boldsymbol{\beta}}_{MLE}=\arg\min\{-l(\boldsymbol{\beta})\}=\arg\min\left\{-\sum_{i=1}^{n}[Y_i\,log\,\pi_i+(1-Y_i)\,log(1-\pi_i)]\right\}$$

最尤法はパラメータ推定において広く使われている方法である．しかし，マイクロアレイデータのように症例数が一般的に 100 例以下であり，独立変数の数（遺伝子数）が数万にも及ぶ“$p\gg n$”の状況では解が定まらず，最尤法を用いることはできない．

2-2.Ridge 推定

“$p\gg n$”における最尤法のかかえる解の不定問題は，回帰係数の L2 ノルム $\left(\sum_{j=1}^{p}\beta_j^2\right)^{1/2}$ の2乗を罰則項に用いたRidge回帰（Hoerl et al. 1988）で対処できる．回帰パラメータの Ridge 推定値 $\hat{\boldsymbol{\beta}}_{ridge}$ は経験ロス関数と罰則項からなる，以下の正則化ロス関数を最小にする解として求まる；

$$\hat{\boldsymbol{\beta}}_{ridge}=\arg\min\left\{-l(\boldsymbol{\beta})+\lambda_2\sum_{j=1}^{p}\beta_j^2\right\}$$

Ridge 回帰は“$p\gg n$”のデータにおいても回帰係数の推定ができるという点では好ましい．しかし，Ridge 推定値 $\hat{\boldsymbol{\beta}}_{ridge}$ は 0 に向かって縮小されるものの，そのほとんどは

正確に 0 とはならない. その結果, マイクロアレイデータ解析から得られるモデルは数万の独立変数からなる複雑なモデルとなってしまう. したがって, モデルに選択された遺伝子の中から応答変数と関連する遺伝子を分子生物学の知識をもとにさらに探索することは容易ではない. つまり, 変数選択の観点から, 実データへの適用には不向きであると考えられる.

2-3.Lasso 推定

罰則項として回帰係数の L1 ノルム $\sum_{j=1}^{p}|\beta_j|$ を利用した回帰分析手法が Lasso である;

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min\left\{-l(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p}|\beta_j|\right\}$$

縮小推定の結果, Ridge 推定値とは異なり Lasso 推定値 $\hat{\boldsymbol{\beta}}_{lasso}$ の多くは正確に 0 となる. したがって, 得られるモデルは少数の独立変数からなる簡素なモデルであり, 変遷選択手法として好ましい性質を持つ.

しかしながら, マイクロアレイデータへの適用を考えるに際して, Lasso には 2 つの大きな欠点があると指摘されている.

1 つは, 独立変数間(遺伝子間)の高い相関を考慮できない手法であるという点である. Lasso では, 相関が高い独立変数から成るグループの中から 1 つの独立変数のみがランダムにモデルに選択されてくる. よって, 真に応答変数と関連がある独立変数をモデルに選択し損なう"偽陰性"が大きくなってしまう. このことは, 同値的に"真陽性"が小さくなってしまうことである.

もう 1 つは, モデルに選択され得る独立変数の数は症例数を超えないという点である. 先に述べたように, マイクロアレイデータは症例数が 100 例以下であることが大半なので, 応答変数と関連のある遺伝子を十分に探索することができない可能性がある.

2-4.Elastic net 推定

独立変数間の高い相関を考慮に入れることができ, かつ, モデルに選択することのできる独立変数の数が症例数に依存しないような, Lasso の欠点を克服した回帰分析手法として, 罰則項に L1 ノルムと L2 ノルムの両方を用いた Elastic net が提案された;

$$\hat{\boldsymbol{\beta}}_{En} = \arg\min\left\{-l(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\right\}$$

Elastic netでは，L1ノルムによる罰則項によって多くの回帰係数の推定値を正確に0に縮小し簡素なモデルを作ることができ，L2ノルムによる罰則項のために相関の高い独立変数をグループとしてモデルに取り込むことができる．そして，選択され得る独立変数の数は最大で p に等しく，症例数に依存しない．したがって，マイクロアレイデータの解析において"真陽性"を大きくできる．

しかし，Elastic net にも欠点があり，それは，"偽陽性"が増大してしまうことである．

2-5.Recursive elastic net 推定

上に述べたように，Lasso と Elastic net は変数選択の視点から比較すると一長一短の関係にある．つまり，Lassoでは，独立変数間の高い相関を考慮できないために"真陽性"が低下してしまう．一方，Elastic netでは，相関を考慮できることから"真陽性"を大きくできるが，Lasso と比べて"偽陽性"が増大してしまう．そこで，相関を考慮できる Elastic net の長所を維持しつつ，"偽陽性"を大きく減少させる手法として提案されたのが Recursive elastic net である．この手法では，Elastic net の L1 ノルムの部分に独立変数毎に異なる重みを加えた罰則項を用いる．そして，事前に設定した任意の繰り返し数 M の範囲で，情報量基準や交差検証法によるモデル評価基準が良くなる限り，回帰係数の推定とその推定値から作られる重みを更新し続ける；

Step1：初期推定値として Elastic net 推定値 $\hat{\beta}_{En} = (\hat{\beta}_{En\_1}, \hat{\beta}_{En\_2}, \cdots\cdots, \hat{\beta}_{En\_p})^{t}$ を求める．

Step2：l 回目において，(l-1)回目の推定値 $\hat{\beta}_{Ren}^{(l-1)} = (\hat{\beta}_{Ren\_1}^{(l-1)}, \hat{\beta}_{Ren\_2}^{(l-1)}, \cdots\cdots, \hat{\beta}_{Ren\_p}^{(l-1)})^{t}$ を用い，独立変数毎の重み $w_j^{(l)} = 1\Big/\left(\hat{\beta}_{Ren\_j}^{(l-1)} + \delta\right)$ を作る $(j = 1, 2, \cdots\cdots, p)$．

Step3：l回目の推定値 $\hat{\beta}_{\mathrm{Ren}}^{(l)} = (\hat{\beta}_{\mathrm{Ren}\_1}^{(l)}, \hat{\beta}_{\mathrm{Ren}\_2}^{(l)}, \cdots\cdots, \hat{\beta}_{\mathrm{Ren}\_p}^{(l)})^{t}$ を以下のように求める；

$$\hat{\beta}_{\mathrm{Ren}}^{(l)} = \arg\min\left\{\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^t\beta)^2 + \lambda_1\sum_{j=1}^{p}w_j^{(l)}\left|\beta_j\right| + \frac{\lambda_2}{2}\sum_{j=1}^{p}\beta_j^2\right\}$$

Step4：中止基準を満たすか，もしくは，任意に設定した最大の繰り返し数 M まで，Step2 と Step3 を繰り返す．ここで，中止基準を満たすとは，モデル評価基準が次の繰り返しで悪くなることを指す．

ここでは応答変数 Y_i は連続値であるので経験ロス関数は残差平方和である．δ は推定値が 0 の時の重みが無限大に発散するのを防ぐためのものであり，経験的に

10^{-5} が妥当であると報告されている. 正則化パラメータである λ_1, λ_2 は各繰り返しの段階で独立に決定される.

Step3 における推定値 $\hat{\boldsymbol{\beta}}_{\mathrm{Re}n}^{(l)}$ は重み付き Elastic net の枠組みにおける最小化問題の解であるが, 重みを使用して独立変数をスケーリングすることで, Elastic net の枠組みでの最小化問題に帰着できることが報告されている;

Step3-1 : 独立変数 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots\cdots, X_{ip})^{\mathrm{t}}$ を重み $\boldsymbol{w}^{(l)} = (w_1^{(l)}, w_2^{(l)}, \cdots\cdots, w_p^{(l)})^{\mathrm{t}}$ で

スケーリングしたものを新たな独立変数 $\boldsymbol{X}_i^* = (X_{i1}^*, X_{i2}^*, \cdots\cdots, X_{ip}^*)^{\mathrm{t}}$ とする;

$$\boldsymbol{X}_i^* = \left(X_{i1} \Big/ w_1^{(l)}, X_{i2} \Big/ w_2^{(l)}, \cdots\cdots, X_{ip} \Big/ w_p^{(l)} \right)^t$$

Step3-2 : Elastic net の枠組みでの解 $\hat{\boldsymbol{\beta}}^{(l)*} = (\hat{\beta}_1^{(l)*}, \hat{\beta}_2^{(l)*}, \cdots\cdots, \hat{\beta}_p^{(l)*})$ を求める;

$$\hat{\boldsymbol{\beta}}^{(l)*} = \arg\min \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^{*t} \boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 \right\}$$

Step3-3 : l 回目の推定値 $\hat{\boldsymbol{\beta}}_{Ren}^{(l)} = (\hat{\beta}_{Ren\_1}^{(l)}, \hat{\beta}_{Ren\_2}^{(l)}, \cdots\cdots, \hat{\beta}_{Ren\_p}^{(l)})^{\mathrm{t}}$ を求める;

$$\hat{\boldsymbol{\beta}}_{Ren}^{(l)} = \left(\hat{\beta}_1^{(l)*} \Big/ w_1^{(l)}, \hat{\beta}_2^{(l)*} \Big/ w_2^{(l)}, \cdots\cdots, \hat{\beta}_p^{(l)*} \Big/ w_p^{(l)} \right)^t$$

そして, 中止基準に従って繰り返しが止まった時の推定値, あるいは, 最後の M 回まで繰り返しが続いた場合には M 回目の推定値が Recursive elastic net 推定値 $\hat{\boldsymbol{\beta}}_{\mathrm{Re}n}$ となる.

3.シミュレーション

3-1.ロジスティック回帰モデルへの導入

応答変数が連続値であることを仮定した Recursive elastic net を応答変数が 2 値の場合にも扱えるようにするため, ロジスティック回帰モデルに Recursive elastic net を拡張することを考えた時, 以下の最小化問題を解く必要が生じ, 計算が複雑になる;

$$\hat{\boldsymbol{\beta}}^{(1)}_{\mathrm{Re}n} = \arg\min\left\{-l(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} w_j^{(1)} \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\right\}$$

そこで，本研究では，Recursive elastic net のロジスティック回帰モデルへの純粋な拡張ではなく，序論で述べた Recursive elastic net の 2 つの考え方をロジスティック回帰モデルへ導入することを考える．

以下に検討手法のアルゴリズムを示す．

Step1：初期推定値として Elastic net 推定値 $\hat{\boldsymbol{\beta}}_{En} = (\hat{\beta}_{En\_1}, \hat{\beta}_{En\_2}, \cdots\cdots, \hat{\beta}_{En\_p})^{\mathrm{t}}$ を求める；

$$\hat{\boldsymbol{\beta}}_{En} = \arg\min\left\{-l(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\right\}$$

Step2：繰り返し数 M の範囲内の l 回目の繰り返し段階において，(l-1)回目の推定値 $\hat{\boldsymbol{\beta}}^{(l-1)}_{Ms} = (\hat{\beta}^{(l-1)}_{Ms\_1}, \hat{\beta}^{(l-1)}_{Ms\_2}, \cdots\cdots, \hat{\beta}^{(l-1)}_{Ms\_p})^{\mathrm{t}}$ を用いることで，独立変数毎に異なる重み $w_j^{(l)} = 1 \Big/ \left(\hat{\beta}^{(l-1)}_{Ms\_j} + \delta\right)$ を作る $(j = 1, 2, \cdots\cdots, p)$．

（ただし，繰り返し数 M は初めに設定する任意の自然数，δ は重み $w_j^{(l)}$ が無限大になるのを避けるための微小な正の値であり，先行研究にならい 10^{-5} と設定した．）

Step3：独立変数の相対的な重要性を解析に考慮できるようにするために重み $\boldsymbol{w}^{(l)} = (w_1^{(l)}, w_2^{(l)}, \cdots\cdots, w_p^{(l)})^{\mathrm{t}}$ で独立変数をスケーリングする．

$$\boldsymbol{X}_i^* = \left(X_{i1} \Big/ w_1^{(l)}, X_{i2} \Big/ w_2^{(l)}, \cdots\cdots, X_{ip} \Big/ w_p^{(l)}\right)$$

ここで，独立変数 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots\cdots, X_{ip})^{\mathrm{t}}$ を重み $\boldsymbol{w}^{(l)} = (w_1^{(l)}, w_2^{(l)}, \cdots\cdots, w_p^{(l)})^{\mathrm{t}}$ でスケーリングした新たな独立変数を $\boldsymbol{X}_i^* = (X_{i1}^*, X_{i2}^*, \cdots\cdots, X_{ip}^*)^{\mathrm{t}}$ と表すものとする．

このスケーリングにより，1 つ前の繰り返し段階での回帰係数の推定値が 0 もしくは 0 に近い微小な値であった独立変数に対しては大きな重みを与えることになる．

これにより，次の推定におけるスケーリング後の独立変数のケース群とコントロール群との差は相対的に小さくなり，次の推定値も0となりやすい．一方，1つ前の繰り返し段階での推定値が0ではない相対的に大きな値となった独立変数には，小さな重み付けをすることになる．これにより，次の推定におけるスケーリング後の独立変数のケース群とコントロール群との差は相対的に大きくなり，次の推定値も0ではない推定値が得られることになる．

Step4：l回目の推定値 $\hat{\beta}_{Ms}^{(l)} = (\hat{\beta}_{Ms\_1}^{(l)}, \hat{\beta}_{Ms\_2}^{(l)}, \cdots\cdots, \hat{\beta}_{Ms\_p}^{(l)})^t$ を求めるにあたり，スケーリング後の独立変数 $\boldsymbol{X}_i^* = (X_{i1}^*, X_{i2}^*, \cdots\cdots, X_{ip}^*)^t$ をもとにした対数尤度関数 $l^*(\boldsymbol{\beta})$ を用いて以下の最小化問題を解き，推定値をさらに重みでスケーリングする；

$$\hat{\boldsymbol{\beta}}^{(l)*} = \arg\min\left\{-l^*(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\right\}$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{Re}n}^{(l)} = \left(\hat{\beta}_1^{(l)*} / w_1^{(l)}, \hat{\beta}_2^{(l)*} / w_2^{(l)}, \cdots\cdots, \hat{\beta}_p^{(l)*} / w_p^{(l)}\right)^t$$

Step5：交差検証法からなる中止基準を満たすか，もしくは，中止基準を満たさない場合にはあらかじめ任意に設定した最大の繰り返し数 M まで，Step2，Step3，Step4を繰り返す．

そして，中止基準に従って繰り返しが止まった時の推定値，あるいは，最後のM回まで繰り返しが続いた場合にはM回目の推定値を検討手法の推定値 $\hat{\boldsymbol{\beta}}_{\mathrm{Re}n}$ とする．

Step1 及び Step4 における推定値を求めるアルゴリズムは，Coordinate Descent(Hastie et al. (2009))[4]や gradient ascent algorithm(Goeman (2009))[8]が近年提案されている．本研究では gradient ascent algorithm を使った R の “penalized”パッケージを用いた．

本研究ではモデル評価基準として 5 分割交差検証法による交差検証法対数尤度を用いた．また，正則化パラメータ λ_2 の探索範囲については，先行研究を参考にしながら，計算コストを考慮して 0.1, 0.5, 1, 2, 5, 10 の 6 点を設定した．λ_1 については，0.05の最小値からすべての回帰係数が0に縮小される時の値を最大値として，細かく探索を行った．

3-2.シミュレーション設定

多次元データにおけるロジスティック回帰モデルでの変数選択に関する先行研究

結果との比較を容易にする目的で，シミュレーションの設定状況は Huang et al. (2008)[5]と同じ設定を考えた．すなわち，ケース群($Y=1$)を 50 例，コントロール群($Y=0$)を 50 例，独立変数の数は 500 個とした．また，応答変数と関連のある独立変数の数は 20 個とした．ケース群とコントロール群の独立変数はそれぞれ以下の 500 次元の多次元正規分布から発生させた；

$$X_i \sim N_{500}(\mu, \Sigma)$$

ケース群：$\mu = (\underbrace{m, m, \cdots\cdots, m}_{20個}, \underbrace{0, 0, \cdots\cdots, 0}_{480個})$

コントロール群：$\mu = (\underbrace{0, 0, \cdots\cdots, 0, 0, 0, \cdots\cdots, 0}_{500個})$

$$\Sigma = \begin{pmatrix} 1 & \rho^{|1-2|} & \rho^{|1-3|} & \cdots & \rho^{|1-500|} \\ \rho^{|2-1|} & 1 & \rho^{|2-3|} & \cdots & \rho^{|2-500|} \\ \rho^{|3-1|} & \rho^{|3-2|} & 1 & \cdots & \rho^{|3-500|} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ \rho^{|500-1|} & \rho^{|500-2|} & \rho^{|500-3|} & \cdots & 1 \end{pmatrix}$$

具体的なm，ρの値は，強シグナルとして$m=1$，弱シグナルとして$m=0.5$，無相関として$\rho=0$，低相関として$\rho=0.3$，高相関として$\rho=0.5$を設定して合計 6 つの組み合わせによる状況を考えた．

上述のデータを200回発生させ，3つの手法の“真陽性”，“偽陽性”を指標にして比較した．

3-3.シミュレーション結果

シミュレーションを 200 回行い，各手法を適用した時の“真陽性”と“偽陽性”の 4 分位点を表 1 にまとめた．

本研究における設定において，繰り返し数 M=1 の検討手法での“真陽性”は Elastic net と比べ減少するものの Lasso よりは大きく，また，“偽陽性”は Elastic net と比較して大きく減少していた．

一方，繰り返し数 M=2 の検討手法における“真陽性”は，状況設定に依らず Lasso と変わらない大きさまで減少していた．“偽陽性”に関しては，M=1 の時より減少するが，シグナルが弱い場合には Lasso より大きく，シグナルが強い場合には Lasso より小さくなった．

表 1　200 回のシミュレーションにおける中央値，25%点，75%点
（応答変数と関連のある独立変数が 20 個，関連のない独立変数が 480 個）

	Lasso		Elastic net		検討手法 (M=1)		検討手法 (M=2)	
n=100	真陽性	偽陽性	真陽性	偽陽性	真陽性	偽陽性	真陽性	偽陽性
μ=1 , ρ=0	18 (16 , 18)	20 (17 , 24)	20 (20 , 20)	90 (61 , 98)	19 (18 , 20)	20 (14 , 25)	17 (16 , 19)	7 (4 , 14)
μ=1 , ρ=0.3	15 (14 , 16)	21 (17 , 24)	20 (18 , 20)	90 (40 , 111)	18 (16 , 19)	32 (21 , 40)	15 (13 , 17)	16 (11 , 23)
μ=1 , ρ=0.5	13 (11 , 14)	21 (17 , 26)	18 (16 , 20)	67 (36 , 111)	16 (13 , 18)	33 (24 , 50)	13 (11 , 15)	18 (13 , 25)
μ=0.5 , ρ=0	12 (10 , 13)	22 (16 , 28)	17 (14 , 19)	94 (39 , 169)	15 (13 , 17)	53 (31 , 83)	13 (11 , 14)	30 (20 , 42)
μ=0.5 , ρ=0.3	9 (7 , 11)	19 (10 , 26)	13 (10 , 17)	52 (27 , 129)	12 (9 , 15)	38 (22 , 75)	10 (8 , 12)	26 (16 , 42)
μ=0.5 , ρ=0.5	7 (5 , 9)	16 (9 , 23)	11 (7 , 15)	35 (17 , 92)	9 (7 , 13)	27 (15 , 59)	8 (6 , 10)	21 (13 , 36)

4．実データへの適用

新たな創薬ターゲットを見つけるための1つの手段として，統計手法を用いて変数選択を行い，次に，分子生物学の知識を利用して，選択されてきた変数の中から重要な予後因子を絞っていくことも可能であると考えられる．その際，統計手法を適用して選択された変数の数は，後の探索における効率性の観点からも重要と思われる．そこで，Lasso，Elastic net，及び，繰り返し数 M=1 の検討手法それぞれを実データに適用した際に選択されてくる変数の数を調べるために，van 't Veer et al. (2002)[7]で扱われたマイクロアレイデータを使用した．

4-1.データの概要

診断時にリンパ節無症状であった 55 歳以下の原発性乳癌患者の腫瘍検体からの遺伝子発情報を利用して，5 年以内に遠隔転移が生じる"予後不良群"(応答変数のコード:1)と5年以内には遠隔転移が生じない"予後良好群"(応答変数のコード:0)を予測することを主な目的に取られたデータである．24481 遺伝子に関する発現情報(log_{10}(ratio))が記録されている．トレーニングデータとして 78 症例あり，そのうち，"予後不良群"が 34 例，"予後良好群"が 44 例である．また，トレーニングデータとは独立したバリデーションデータが 19 症例分ある．19 症例のうち，"予後不良群"が 12 例，"予後良好群"が 7 例である．

4-2.解析方法

本研究の比較指標は予測ではなく変数選択であるので，トレーニングデータとバリデーションデータとを分けず，97 症例を解析に用いた．また，本研究におけるシミュレーションで検討したのは独立変数の数が 500 個の状況であった．よって，実データへの適用においても独立変数の数を同様に 500 個とするため，以下のスクリーニングを行った．

Step1:発現情報に欠測が多いと妥当な結果を導くことができないので，30%以上の症例に欠測がみられる遺伝子を除いた．

Step2:欠測症例が 30%未満の遺伝子における欠測には，欠測ではない症例の中央値で補完した．

Step3:平均 0，分散 1 となるように発現情報を標準化した．

Step4:各遺伝子の発現情報と 2 値の応答変数の単相関係数(スピアマンの順位相関係数)を算出し，その絶対値が大きい方から順に 500 個の遺伝子を選択した．

スクリーニングにより選択された 500 個の遺伝子データに対して，3 つの手法を適用し，回帰係数の推定値が 0 ではない独立変数(遺伝子)の数を求めた．Elastic net と検討手法の正則化パラメータ λ_2 は，シミュレーションと同様に 0.1，0.5，1，2，5，10 の 6 点に設定し，

もう1つの正則化パラメータλ_1とともに，モデル評価基準に5分割交差検証法による交差検証法対数尤度を用いて最適な値を決定した．

4-3.結果

各手法を適用して選択された遺伝子の数はLassoで31個，Elastic netで292個，検討手法で97個となった．Elastic netにおいて選択された遺伝子数はLassoや検討手法と比較して多いものであった．そして，その多くは“偽陽性”であると推察されるので，以後の関連遺伝子の探索効率は低いであろう．しかしながら，実データであるために選択されたもののうち，どの遺伝子が“真陽性”あるいは“偽陽性”であるのかは明らかではない．選択された遺伝子がどのようなグループを構成しながら応答変数に寄与しているのかを視覚的にとらえることができれば，関連遺伝子の分子生物学的観点からのさらなる探索において有効と思われる．そこで，遺伝子間の距離をユークリッド距離で定義し，多次元尺度構成法を用いて遺伝子間の近さを2次元平面に配置したものを図1～図3に表す．Elastic netでは，選択された遺伝子が多すぎるためデータ構造がわかりにくい．また，Lassoでは選択された遺伝子が少数であることから遺伝子のグループ情報が把握しにくい．これらに対して，検討手法では構成している遺伝子グループが判断しやすくなっている．つまり，遺伝子のグループ情報を解析に考慮できていることがうかがわれる．

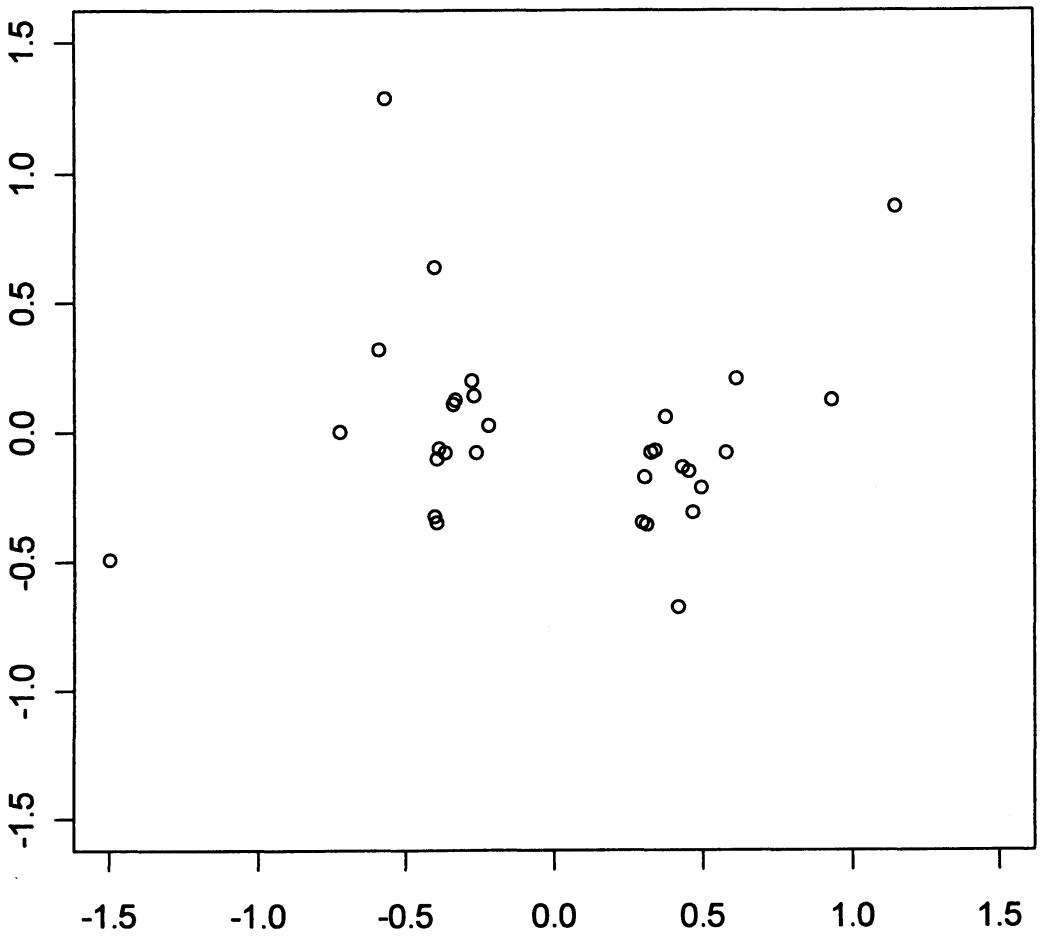

図1 多次元尺度構成法(Lasso,31遺伝子)

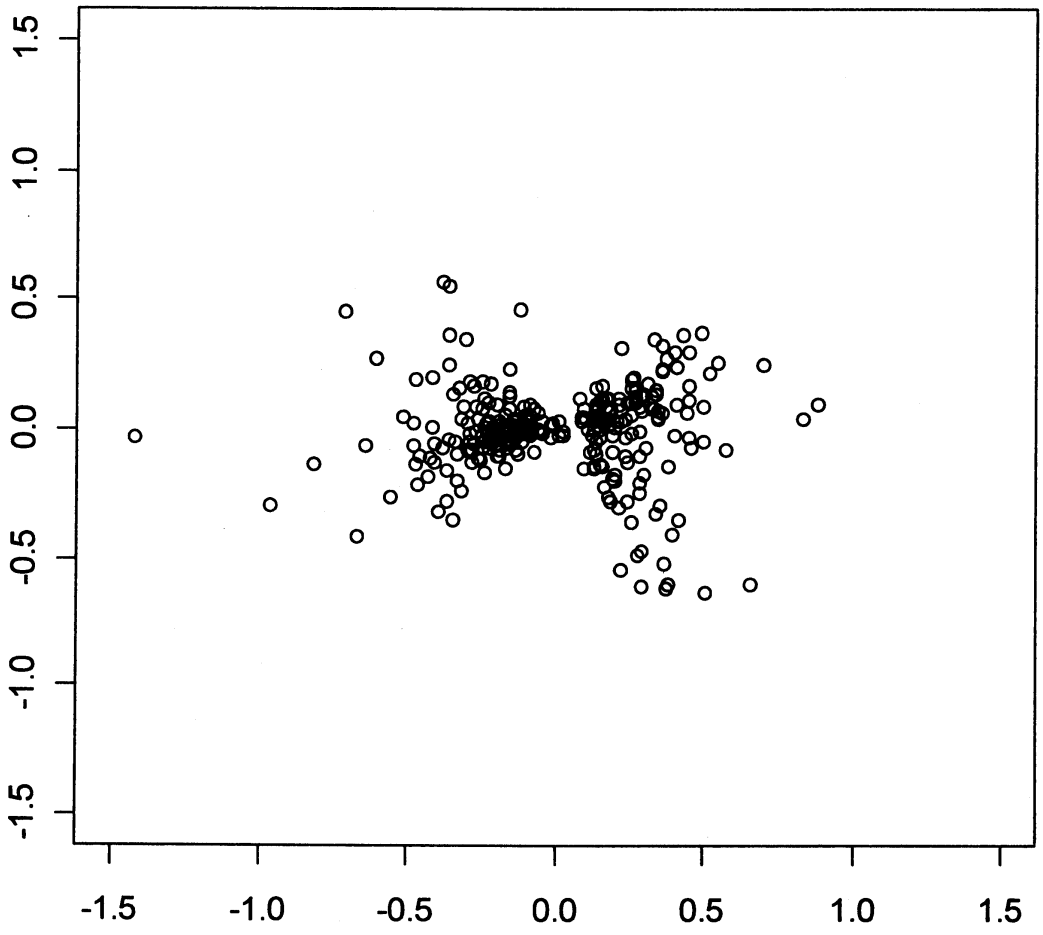

図2 多次元尺度構成法(Elastic net,292遺伝子)

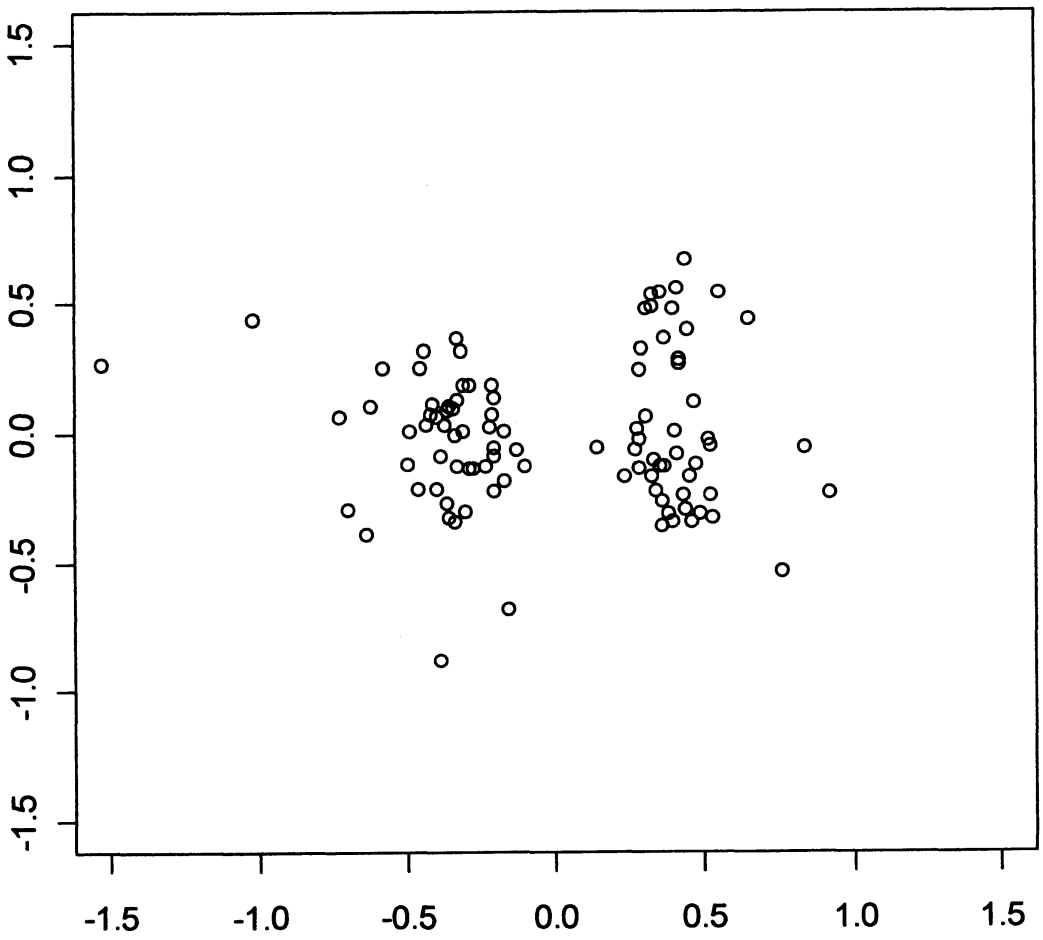

図3 多次元尺度構成法(検討手法,97遺伝子)

5.考察

現在，Shimamura et al.の提案した手法である Recursive elastic net は応答変数に連続値を仮定しており，応答変数が 2 値の場合にも扱えるようにロジスティック回帰モデルに応用した統計手法の性能は未知である．したがって，本研究では，Recursive elastic net のもつ，(1)変数独立変数毎の相対的重要性を考慮できる，(2)繰り返しの推定を行う，というアイデアをロジスティック回帰モデルへ導入した検討手法を変数選択の点から，Lasso，Elastic net と比較検討した．

今回のシミュレーション結果から，以下の 2 点が示唆された．1 つ目は，繰り返し数 M は 1 回で十分ということである．2 回の繰り返しを行うと“偽陽性”の減少という点では好ましい性質を持つ半面，“真陽性”が Lasso と同等まで低下してしまった．この M=2 における結果を，初期推定値として Lasso 推定値を利用した adaptive lasso の枠組みで捉えることのできる統計手法である Iterated lasso(Zhang et al. (2008))[5]の結果と比較すると，Iterated lasso の方が“真陽性”，“偽陽性”の点で優れていた．また，応答変数が連続値の場合には 2 回以上の繰り返しが最適な場合もあり得るが，“偽陽性”は 1 回目の繰り返しによって大きく減少し，2 回目以降の繰り返しでみられる“偽陽性”の減少はわずかであることが報告されている．このことからも繰り返し数を 1 回とすることは妥当であると考えられる．

2 つ目は，繰り返し数 1 回の検討手法は，応答変数が連続値の Recursive elastic net と類似した性質を持つであろうということである．つまり，M=1 の検討手法は，Elastic net の欠点である大きな“偽陽性”を大幅に削減でき，かつ，“真陽性”に関しては Elastic net よりは低下するものの，独立変数間の高い相関関係を考慮できることにより，Lasso より大きいものとなった．

今回のシミュレーションは，先行研究結果[5]との比較を簡単にするために当該先行研究と同じ設定の下で行った．しかしながら，当然，各手法における“真陽性”の大きさや差の程度は，シミュレーション設定によって変わってくる．そこで，より日本での臨床研究を反映できるような設定で新たなシミュレーションを行い，結果を表 2 に示した(シミュレーション設定の詳細，及び結果については appendix1)．

新たなシミュレーション結果からも，検討手法の性質に関して，

(1)繰り返し数 M は 1 回で十分である．

(2)M=1 での検討手法の“真陽性”は Lasso よりは大きく Elastic net よりは小さくなる傾向にある．

(3)M=1 での検討手法の“偽陽性”は Elastic net より大きく削減できる．

という，先のシミュレーション結果と共通する示唆が得られた．また，表 2 からわかるように，状況によっては，M=1 の検討手法によって Lasso と変わらない大きさまで “偽陽性”を減らせる得ることがわかった．

実データへの適用に際して，どの手法を用いるべきかという問題は，有益な情報の増加

と無駄な情報の削減のどちらに重きを置くかということに依存して決まるものであろう. ただ, Elastic net は“偽陽性”が他の 2 手法と比べて大きいことで, {“真陽性”/ (“真陽性”+“偽陽性”)}で定義される“正確率”が小さくなるので, 関連遺伝子探索には不向きといえよう. また, Lasso によって選択できる独立変数の数が症例数を超えないという性質を考慮すると, 症例数の集まりにくい日本での臨床研究を想定した場合, Lasso は実データ解析に使いにくいと思われる. 一方, M=1 の検討手法は{“真陽性”/ (“真陽性”+“偽陰性”)}で定義される“感度” と“正確度”の両者のバランスをうまく勘案した手法に位置付けられるものと考えられる. さらに, 表 1 と表 2 の結果から, 真に応答変数と関連のある独立変数の数が大きい程, 各手法における“真陽性”の差は大きくなることが窺われた. 同じ生物学的経路を共有する遺伝子は高い相関を示す[3]とすると, 応答変数と関連のある独立変数の数は 50 程度の症例数よりも大きい状況もあると思われる. この場合, Lasso と M=1 の検討手法の“真陽性”の差は大きくなり, 変数選択の点での M=1 の検討手法の好ましさが一層際立つと考えられる.

正則化パラメータの最適な値は応答変数と真に関連のある独立変数の数, 症例数, 独立変数のシグナルの大きさ, そして, 独立変数の分散共分散構造などに依存して変わってくるであろう. しかし, 本研究では計算コストを考慮して, 正則化パラメータ λ_2 の探索はシミュレーション設定や繰り返しの段階に依らず常に同一の 6 点とした. 状況によっては, この 6 点が全く見当外れの範囲を探索している可能性もあり, そのことで“真陽性”及び“偽陽性”の結果に大きく影響してきていることも考えられる. この点は本研究の限界の 1 つである。

変数選択において望ましい性質である oracle property [11]をもつ手法として, adaptive Elastic-Net(Zou et al. (2009))[10]が正規線形モデルの枠組みにおいて本研究中に提案された. この手法は繰り返し数 M=1 の検討手法と非常に類似している. 重みの構成の仕方と正則化パラメータ λ_2 を Elastic net 推定値を求める最初の段階で決定された値で固定している点が異なっている. λ_2 を各繰り返し段階で独立に決定する必要がなければ計算の負担は大きく軽減されるうえに, oracle property との関連からも λ_2 を初めの値に固定したまま推定と重みの更新を繰り返した時に今回の結果よりも良い結果が得られる可能性があり, 今後検討する価値はあると思う.

また, 初期推定値や重みに何を用いるかでも結果は変わってくると推測される. この初期推定値や重みに本研究とは異なるものを用いることで, Elastic net の“真陽性”をほとんど下げることなく“偽陽性”を大きく減少できる統計手法となり得るかもしれず, 今後の課題としたい.

Appendix1

・シミュレーション設定

日本での臨床研究では症例数を 100 例も集めることは難しい. そこで, ケース群 ($Y=1$) を 25 例, コントロール群($Y=0$)を 25 例, 独立変数の数は 500 個とした. また, 関連遺伝子

数が症例数を上回る状況を想定し，応答変数と関連のある独立変数の数は 80 個とした．ケース群とコントロール群の独立変数はそれぞれ以下のような 500 次元の多次元正規分布から発生させた；

$$\boldsymbol{X}_i \sim \boldsymbol{N}_{500}(\mu, \Sigma)$$

ケース群：$\mu = (\underbrace{\boldsymbol{m}, \boldsymbol{m}, \cdots\cdots, \boldsymbol{m}}_{80個}, \underbrace{\boldsymbol{0}, \boldsymbol{0}, \cdots\cdots, \boldsymbol{0}}_{420個})$

コントロール群：$\mu = (\underbrace{\boldsymbol{0}, \boldsymbol{0}, \cdots\cdots, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}, \cdots\cdots, \boldsymbol{0}}_{500個})$

$$\Sigma = \begin{pmatrix} \mathbf{1} & \rho^{|1-2|} & \rho^{|1-3|} & \cdots & \rho^{|1-500|} \\ \rho^{|2-1|} & \mathbf{1} & \rho^{|2-3|} & \cdots & \rho^{|2-500|} \\ \rho^{|3-1|} & \rho^{|3-2|} & \mathbf{1} & \cdots & \rho^{|3-500|} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ \rho^{|500-1|} & \rho^{|500-2|} & \rho^{|500-3|} & \cdots & \mathbf{1} \end{pmatrix}$$

具体的な m，ρ の値は，強シグナルとして $m=1$，無相関として $\rho=0$ ，低相関として $\rho=0.3$ ，高相関として $\rho=0.5$ 設定して合計 3 つの組み合わせによる状況を考えた．

計算コストを考慮し，正則化パラメータ λ_2 の探索は 0.1, 0.5, 1, 2, 5, 10 の 6 点とした．正則化パラメータ λ_1 の探索は，すべての縮小推定値を 0 とする最大値から 0.1 を最小値として細かく行った．そして，最適な値は 5 分割交差検証法による交差検証法対数尤度を用いて決定した．結果は表 2 に示した．

表 2　200 回のシミュレーションにおける中央値, 25%点, 75%点
（応答変数と関連のある独立変数が 80 個, 関連のない独立変数が 420 個）

	Lasso		Elastic net		検討手法 (M=1)		検討手法 (M=2)	
n=50	真陽性	偽陽性	真陽性	偽陽性	真陽性	偽陽性	真陽性	偽陽性
μ=1 , ρ=0	27 (25 , 29)	1 (0 , 1)	74 (70 , 75)	36 (25 , 40)	74 (70 , 75)	35 (26 , 40)	74 (70 , 75)	35 (26 , 40)
μ=1 , ρ=0.3	25 (23 , 27)	2 (1 , 3)	70 (53 , 73)	44 (12 , 51)	42 (40 , 58)	5 (3 , 13)	69 (19 , 73)	39 (1 , 48)
μ=1 , ρ=0.5	23 (21 , 24)	3 (2 , 5)	70 (66 , 73)	56 (41 , 65)	43 (39 , 76)	9 (5 , 18)	22 (18 , 70)	2 (0 , 49)

6.参考文献

[1] Tibshirani Robert. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58:267-288.

[2]Zou Hui, Hastie Trevor. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301-320.

[3] Mark R. Segal, Kam D. Dahlquist, Bruce R.Conklin. (2003). Regression approach for microarray data analysis. *Journal of Computational Biology* 10(6): 961–980.

[4] Friedman Jerome, Hastie Trevor, Tibshirani Robert. (2010). Regularized Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1),

[5] Jian Huang, Shuange Ma, Cun-Hui Zhang. (2008). The Iterated Lasso for High-Dimensional Logistic Regression. The University of Iowa Department of Statistical and Actuarial Science Technical Report No.392

[6] Shimamura Teppei, Imoto Seiya, Yamaguchi Rui, Fujita André, Nagasaki Masao, Miyano Satoru. (2009). Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology* **3**:41

[7] Laura J. van 't Veer, Hongyue Dai, Marc J. van de Vijver,Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, ReneÂ Bernards, Stephen H. Friend. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530-536.

[8] Jelle J.Goeman. (2009). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 51(6):1–15

[9] Zou Hui. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association* 101(476):1418-1429.

[10] Zou Hui, Hao Helen Zhang. (2009). On the adaptive elastic net with a diverging number of parameters. *The Annals of Statistics* 37(4):1733-1751

[11] Jianqing Fan, Runze Li. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* 96(456): 1348-1360